공학석사학위논문

조건부 변분 자동인코더 기반 변환 음성의 억양 다양화 방법 연구

Voice Conversion with Diverse Intonation using Conditional Variational Auto-Encoder

2018 년 8 월

서울대학교 대학원

산업공학과

서 수 빈

조건부 변분 자동인코더 기반 변환 음성의 억양 다양화 방법 연구

Voice Conversion with Diverse Intonation using Conditional Variational Auto-Encoder

지도교수 박 종 헌

이 논문을 공학석사 학위논문으로 제출함

2018 년 6 월

서울대학교 대학원

산업공학과

서 수 빈

서수빈의 공학석사 학위논문을 인준함

2018 년 6 월

위 원 장 조 성 준 (인)

부위원장 박 종 헌 (인)

위 원 이 재 욱 (인)

초록

음성 변환(Voice Conversion)은 원천 화자의 언어 정보를 유지하면서 변환 대상 화자의 음성과 발화를 합성하는 작업이다. 즉, 목소리 데이터에서 언어적 특징과 스타일 특징을 분리할 수 있다는 가정하에, 스타일 특징만을 변환하는 작업을 일컫는다.

기존에 제시된 음성 변환 연구들은 스타일 변환을 하나의 함수 형태로 모델링하여 결정론적이다. 즉, 기존의 음성 변환 모델은 한 원천 화자의 입력에 대해 단 하나의 억양을 가지는 발화만을 생성하는 것으로 제한된다. 하지만 실제 화자는 하나의 스크립트에서 다양한 억양을 가지는 발화를 생성 할 수 있기 때문에 기존 연구들은 한계점을 지닌다.

이러한 한계를 극복하기 위해 모델에 확률적인 정보를 부여하기 위한 방법으로 변분 자동 인코더(Variational Auto-Encoder)를 사용하였다. 변분 자동 인코더는 딥러닝의 생성형 모델의 일환으로, 확률분포에서 잡음을 샘플링하여 기존 데이터의 분포를 생성해내는 모델이다. 본 논문에서는 화자의 억양, 즉 스타일을 담당하는 부분을 변분 자동 인코더로 모델링함으로써 분포에서 샘플링 할 때 마다 다양한 억양을 가지는 음성을 합성하는 새로운 접근법을 제안한다.

실험에 따르면 화자의 스타일 기능을 가우시안 분포의 잠복 공간으로 매핑 할 수 있다는 것이 입증되었다. 특히, 2-3 시그마 범위 내에 존재하는 샘플들에 대해 억양이 확연하게 바뀌는 것을 확인하였다. 본 논문은 또한 잠복 공간의 사후 분포를 역 자기 회귀성 유동 (Inverse Autoregressive Flow)로 보다 복잡하게 만드는 방법을 제시함으로써 보다 다양한 억양으로 음성을 변환 할 수 있었다. 결과적으로 변환된 음성은 다양한 억양을 가질 뿐만 아니라 기존의 결정론적 모델보다 우수한 음질을 제공한다.

주요어: 음성 변환, 변분 자동인코더, 억양

학번: 2016-24163

목차

초록 **i**

목차 **v**

표 목차 **vi**

그림 목차 **vii**

제 1 장 서론 **1**

1.1 연구 배경 1

1.2 연구 내용 및 공헌 4

1.3 논문구성 6

제 2 장 배경이론 및 관련 연구 **7**

2.1 음성신호처리 7

2.1.1 이산화 및 양자화 7

2.1.2 스펙트로 그램 변환 7

2.1.3 보코더 8

2.2 음성 변환 10

2.3 변분 자동인코더 12

2.3.1 목적 함수 12

2.3.2 최적화 13

2.3.3 조건부 변분 자동인코더 14
2.4 역 자기회귀성 유동 15
2.5 관련 연구 18

제 3 장 제안 기법 **20**

3.1 음소 분류기 20
3.2 음성 합성기 22
3.3 샘플 생성 24
3.4 역 자기회귀성 유동 적용 25
3.5 모델 구조 27

제 4 장 실험 및 결과 **29**

4.1 실험 설계 29
4.2 음소 분류기 학습 결과 32
4.3 다양한 억양 34
4.4 품질 평균 의견 점수 (MOS quality) 37
4.5 절제 연구 40
4.5.1 역 자기회귀성 유동의 효과 40
4.5.2 멜 스펙트로그램과 선형 스펙트로그램의 차이 40
4.5.3 프리빗 알고리즘 42

제 5 장 결론 **45**

5.1 결론 45
5.2 향후 발전 방향 47

참고문헌 **49**

Abstract **54**

표 목차

표 4.1 음성데이터 전처리 하이퍼-파라미터 30
표 4.2 Mean Opinion Score (MOS) 37

그림 목차

그림 2.1 음성 변환 도식화 10
그림 2.2 재 매개변수화 속임수 [1] 14
그림 2.3 역 자기회귀성 유동의 효과 [2] 15

그림 3.1 전체 모델 구조 28

그림 4.1 PPGs 샘플 32
그림 4.2 ϵ에 따른 억양의 변화 35
그림 4.3 ϵ에 따른 억양의 변화 2 36
그림 4.4 10000번 훈련 결과 38
그림 4.5 30000번 훈련 결과 39
그림 4.6 50000번 훈련 결과 39
그림 4.7 역 자기회귀성 유동 적용 비교 실험 결과 41
그림 4.8 $\lambda = 1$일 때 멜-스펙트로그램 변화 43
그림 4.9 $\lambda = 2$일 때 멜-스펙트로그램 변화 44

제 1 장 서론

1.1 연구 배경

다른 사람들, 특히 유명인들의 목소리를 모방하는 것은 끊임없이 관심을 끄는 주제다. 이에, 다른 사람들의 목소리를 모방하는 방법으로 문자를 특정인의 목소리로 읽어내는 문자-음성 변환 방식 (Text-to-Speech)과, 기존 화자의 음성에서 특정 화자의 음성으로 목소리를 바꾸는 음성 변환 방식 (Voice Conversion)이 대두되었다. 두 가지 방법은 그 파이프라인이 거의 동일하다. 음성 변환 방식은 문자-음성 변환 방식의 한 일종으로, 음성-문자-음성의 과정을 거치게 된다. 즉, 원천 화자의 음성에서 문자에 관한 정보를 추출해내고, 이를 다시 목표 화자의 음성으로 생성해내는 것이다. 상대적으로 음성-문자 변환은 문자-음성 변환 방식에 비해 쉬운 편이기 때문에 문자-음성 변환 방식을 잘 모델링하는 것이 음성 변환 문제의 핵심이 되었다. 따라서, 음성 변환 연구들은 문자-음성 변환 방식의 연구에 초점을 맞추고 있다.

기존의 문자-음성 변환 연구의 파이프라인은 문자열과 음성 신호에 대한 전처리 과정 및 다양한 모듈들이 필수적이기 때문에 매우 복잡한 과정을 거치게 된다 [3]. 우선 문자열을 음소 기호로 변환해야 하며, 각 문자열이 발화하는 음성의 시간 길이에 따라 문자열과 음성 신호를 시간대별로 나누어 쌍으로 맞추어 주는 작업이 필요하다. 또한 특정 화자의 주파수를 예측하는 작업이나 샘플링 방식 등 복잡한 신호처리가 뒤따른다. 최근에 이런 복잡한 파이프라인을 심층 신경 네트워크 구조(Deep Neural Network)를 이용하여 하나의 커다란 신경망으로 대체하는 연구가 등장하면서, 문자-음성 변환 방식의 새로운 패러다임을 제시하였다 [4]. 본 연구에서도 복잡한 파이프라인을 사용하는

대신 심층 신경 네트워크 구조를 이용하여 전체 구조를 구성하였다.

심층 신경 네트워크는 이와 같이 다양한 분야에서 유망한 결과를 보여 주었고, 음성 변환 모델에서도 심층 신경 네트워크를 사용하는 모델이 등장하기 시작하였다. 심층 신경 네트워크 학습을 통한 음성 변환 모델은 언어적 기능과 스타일 기능으로 구분할 수 있는 잠재적 공간으로의 매핑을 배우는 것을 목표로 한다. 그런 다음 모델은 기존 화자의 발화의 언어적 정보를 유지하면서 변환 대상의 화자에 대한 스타일 정보를 적용한다. [5] 및 [6]은 동일한 스크립트를 읽는 소스 및 대상 데이터 쌍이 있는 병렬 코퍼스에서 음성 변환 모델을 학습하였다.

위와 같이 쌍으로 존재하는 데이터는 획득하기 어렵고, 크기도 제한되어있기 때문에, 비병렬 데이터에 대해 비감독 방식(Unsupervised Learning)으로 훈련된 음성 변환 모델이 연구되었다. [7]는 화자의 특성을 포함하고 문자열과 독립적인 기본 주파수 (F0) 기능을 사용하여 음성 신호를 변환하였다. 이 방법을 사용하여 다양한 화자 데이터에 대한 모델을 학습한 다음 F0이 화자의 스타일을 대표하는 피쳐라고 가정하고 각 F0에 대해 다른 목소리를 생성한다. 비슷한 방식으로, [8]는 음소의 확률을 나타내는 음소 사후 분포 그램 (Phoneme Posterior Grams, PPGs) 개념을 도입하였다. 기존 화자로부터 추출된 PPG가 화자의 스타일과 무관하다는 가정에서 시작하여, 화자는 화자의 발화로부터의 PPG와 F0를 이용하여 변환 된 음성을 합성한다.

그러나, 이 모델은 학습되어 파라미터들의 고정되어있기 때문에 하나의 기존 화자 발화에 대해 단지 하나의 억양만을 갖는 결정론적인 출력을 갖는다는 한계가 존재한다. 동화를 읽거나 많은 사람들 앞에서 연설을 할 때, 사람들은 똑같은 문장을 읽더라도 매번 다른 억양으로 발화한다. 이것은 다른 사람들의 스타일 차이가 아니라 한 개인의 발화 간에도 차이가 있음을 시사한다. 본 논문에서는 사람이 발화할 때 다양한 억양을 가지는 목소리를 생성하기 위해서는 확률적인 정보가 필요하다는 가정을 한다. [9]에

는 억양을 위한 특징을 추출하려고 시도하지만 신경망을 사용하는 end-to-end 방식은 아니며 확률적 정보는 다루지 않았다.

최근에는 변분 자동 인코더 (Variational Auto-Encoder) 및 생성형 적대 네트워크 (Generative Adversarial Network)과 같은 다양한 확률론적 생성 모델이 도입되어 다양한 생성 문제에 대해 확률 정보를 사용할 수 있게 되었다. [10]와 [11]에서 변분 자동 인코더를 사용하여 언어적 정보를 나타내는 확률적 잠재 변수를 발화로부터 학습한다.

1.2 연구 내용 및 공헌

본 연구에서는 다양한 억양을 가지는 발화를 합성하기 위해 확률적 정보를 변분 자동 인코더를 이용하여 모델에 부여한다. 변분 자동 인코더는 비교적 단순한 가우시안 분포에서 샘플링된 사전확률분포와, 데이터를 조건부로 하는 사후확률분포를 가깝게 하면서 학습하는 방식의 모델이다.

모델은 크게 음성-문자 변환을 담당하는 음소 분류기(Phoneme Classifier)와 변환된 음소를 다시 음성으로 변환하기 위한 문자-음성 변환을 담당하는 음성 생성기(Speech Synthesizer)로 구성된다. 음소 분류기는 원천 화자의 음성에서 스타일을 제외한 문자열의 정보만을 추출해내며 음성 합성기는 조건부 변분 자동 인코더 구조를 이용하여 억양을 나타내는 잠재 공간을 학습하고 언어적 정보는 앞서 예측된 음소 분류기의 출력값으로 고정하여 조건부 변수로 사용한다. 특정 억양을 조건부 변수로 고정하는 것은 쉽지 않으며 스타일에 대해 레이블링된 데이터를 구하기도 쉽지 않으므로 잠재적인 공간에서 무작위 값을 샘플링하는 것은 다양한 억양을 생성하기에 좋은 접근법이다.

또한, 비교적 단순한 가우시안 분포를 사후확률 분포로 사용하는 기존의 변분 자동 인코더와는 달리 본 모델에서는 가우시안 분포에 비해 복잡한 사후확률분포를 사용하기 위하여 역 자기회귀성 유동(Inverse Auto-regressive Flow) 방식을 적용하였다. 이 방식을 적용하면 더 복잡한 사후 확률분포를 사용하기 때문에 기존의 가우시안 분포를 사용하는 변분 자동 인코더에 비해 더 다양한 억양을 가지는 음성을 생성할 수 있을 것이라 기대하였다.

결과적으로, 똑같은 억양을 가지는 발화만을 생성하는 기존의 결정론적 모델과는 달리, 본 모델은 다양한 억양으로 단일 화자의 발화를 생성할 수 있다. 기존의 음성 변환 연구들 중에서 억양을 다양화하려는 시도는 없었으며, 확률적 정보를 사용하여 접하지 않은 데이터에 대해 성능 저하없이 다양한 억양으로 변환된 발화를 만들어 냈다. 또한,

역 자기회귀성 유동 방식을 통해 음성 전체적으로 기존 변분 자동 인코더 방식에 비해 더욱 다양한 억양을 가지는 음성을 생성해내는 것을 확인하였다. 추가적으로, 제안된 모델은 확률적 정보없이 학습된 모델에서 생성된 음성보다 자연스러움에서 더 나은 성능을 보였다.

1.3 논문구성

본 논문은 총 5 장으로 구성된다. 제 2장에서는 음성 변환 및 변분 자동 인코더 관련 선행 연구를 살펴본다. 제 3장에서는 다양한 억양을 가지는 음성 변환에 대한 모델을 제안한다. 제 4장에서는 제시한 실험의 결과를 살펴본다. 마지막으로 제 5장에서는 결론과 향후 연구방향을 제시한다.

제 2 장 배경이론 및 관련 연구

2.1 음성신호처리

2.1.1 이산화 및 양자화

자연적으로 존재하는 음성 신호는 아날로그 형식으로, 연속적인 형태를 띈다. 음성 신호를 컴퓨터가 처리할 수 있는 형태로 바꾸기 위해서는 이산화 및 양자화 과정이 필수적이다. 연속적인 시간의 아날로그 신호를 구간 단위의 시간대로 샘플링하는 방법을 이산화(Discretization)라고 하며, 표본추출 된 신호를 컴퓨터가 인지할 수 있는 비트(bit) 형태로 구간별로 나누어 주는 것을 양자화(Quantization)이라고 한다. 즉, 이산화 과정은 시간축에 대해 구간을 나누는 것이며, 양자화는 진폭축에 대해 구간을 나누는 것이라 할 수 있다. 이산화 과정 시 무작위로 샘플링을 하게 되면 정보의 손실이 나타나게 된다. 정보의 손실 없이 이산화하는 방법은 나이퀴스트 정리에 따라 원본의 최대 주파수의 2배 이상의 표본주파수로 표본을 추출하여야 한다. 나이퀴스트 정리에 따르면, 표본화 주파수가 2배보다 작을 경우, 고주파끼리의 간섭이 일어나 정보가 손실된다. 양자화는 8bit 또는 16bit의 형태로 하며, 8bit는 256개의 구간, 16bit는 65536개의 구간으로 나누어 연속적인 진폭을 구간별로 나눈다. 이산화와 양자화 과정을 모두 마치면, 분석 가능한 디지털화 된 음성 데이터를 얻을 수 있다.

2.1.2 스펙트로 그램 변환

디지털화된 음성 데이터는 그 자체로 분석이 가능하지만 분석하기에 크기가 너무 크고, 특정 피쳐로서의 역할을 하기에 성능이 떨어지기 때문에 대부분의 음성신호 분

석에서 음성 데이터를 스펙트로 그램으로 변형하여 분석한다. 또한, 스펙트로 그램이 시간과 주파수의 2차원 형태의 피쳐로, 마치 그림과 같기 때문에 음성 신호를 시각화하여 분석하는 것에도 유용하게 사용되므로 음성학에서 자주 사용되고 있다. 기존의 음성 데이터를 스펙트로 그램으로 변형하는 방법은 주파수대별로 단시간 푸리에 변환(STFT)을 적용한다. 단시간 푸리에 변환의 식은 다음과 같다.

$$STFTx[n](m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn} \tag{2.1}$$

변환된 스펙트로 그램은 전체 주파수 영역대를 모두 똑같은 가중치로 변형한 형식이다. 본 연구에서는 인간의 발화의 변환에 관한 연구이기 때문에 고주파수대의 음성신호는 상대적으로 불필요하다. 이를 위해 멜-스펙트로 그램(Mel-spectrogram)을 사용한다. 멜-스펙트로그램이란 인간이 들을 수 있는 고유 주파수 영역대에 가중치 가지는 멜-필터(Mel-filter)를 적용하여 스펙트로 그램을 한번 더 변환한 피쳐다. 멜-필터를 사용하게 되면 인간이 들을 수 있는 주파수 영역대인 저주파수의 특징이 더욱 강조되고, 주파수 영역대가 로그-스케일로 늘어나게 된다.

2.1.3 보코더

스펙트로그램은 실제 음성 신호에서 위상값을 제거하였기 때문에 다시 들을 수 있는 형태인 음성 신호로 복구하기 위해서는 위상값을 복원해주는 보코더가 필요하다. 보코더는 음성의 크기값만을 담고있는 스펙트로그램이나 멜-스펙트로그램을 실제 음성 신호 형태로 다시 복원하는 것을 목적으로 한다. 기존의 보코더는 가우시안-혼합모델(Gaussian Mixture Model)을 사용하여 선 스펙트럴 쌍이나 비주기성 파라미터를 예측하여 실제 음성 신호를 복원하였다. 최근에는 [12]에서 제시한 종단 신경망 보코더를 이용하여 오직 신경망만으로도 음성신호의 복원이 가능해졌으며, 기존의 보코더보다

성능도 더 개선됨을 확인하였다. 하지만 구현이 어렵고 학습시간이 매우 오래걸린다는 단점이 존재하며, 웨이브넷의 자기 회귀성 구조 특성 상 음성 신호를 복원하기 위해 매우 오랜 시간이 소요된다는 점에서 한계점이 존재했다. 이를 해결하기 위해 [13]는 역 자기 회귀성 유동을 이용하여 자기 회귀성 구조를 없애면서 빠르게 음성 신호를 복원하는 방법을 고안하였고, 현재 구글 어시스턴트에 탑재되어 사용되고 있다. 이와는 별개로, 본 논문에서 사용한 간단한 보코더 방식은 그리핀-림 알고리즘(Griffin-Lim Algorithm) [14]이다. 이 방식은 반복적인 수행을 통해 복원값에 가장 가까운 역 단시간 푸리에 변환을 찾아내는 방식이다. 이 방법은 간단하면서도 빠른 복원이 가능하기 때문에 현재도 널리 이용되고 있다.

2.2 음성 변환

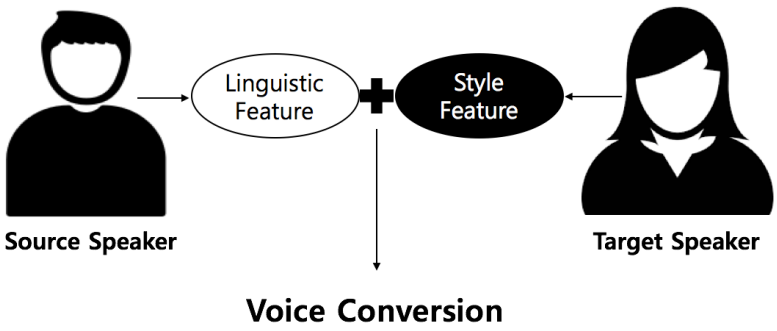

Figure 2.1: 음성 변환 도식화

음성 변환이란 원천 화자의 발화에서 언어적 특징만을 추출해낸 후 목표 화자의 목소리, 즉 스타일을 입혀 원천 화자가 발화한 문장을 목표 화자의 음성으로 발화하는 것을 의미한다. 음성 변환은 음성 데이터에서 언어적 특징과 스타일 특징을 어떻게 분리할 것인가가 핵심이다. 음성 변환은 크게 2가지로 나뉘는데, 동일한 문장을 말하면서 발화 시간이 거의 비슷한 서로 다른 두명의 화자가 발화한 데이터쌍이 존재할 경우, 그리고 위와 같은 데이터쌍이 존재하지 않을 경우 두가지로 나뉜다. 첫번째의 경우, 언어적 특징이 완전히 동일하기 때문에 언어적 특징에 대해 고려할 필요 없이 스타일에 대한 변환만 고려하면 된다. 딥러닝을 이용하면 각 프레임별로 목표 화자에 가깝게 학습하게 되면 비교적 쉽게 학습이 가능하다.

하지만 완전히 동일한 문장에 대해 비슷한 발화 시간을 가지는 데이터쌍이 매우 부족하며, 다양한 목표화자로의 음성 변환이 어렵기 때문에 이 방법은 선호되지 않는다. 따라서 두번째로 제시된 데이터 쌍이 필요없는 음성 변환에 대한 연구가 활발히 진행되었다. 이 방법은 음성 데이터에서 음성 인식을 통하여 언어적 정보를 추출한 뒤,

목표 화자에서 화자의 목소리 스타일 특성을 가장 잘 나타내는 F0 피쳐를 뽑아 다시 음성을 합성하는 형태였다.

최근 딥러닝에서 다양한 생성 모델이 조명을 받으면서 적대적 생성 네트워크나 변분 자동 인코더를 이용한 음성 변환에 관한 연구도 활발히 진행되었다. 적대적 생성 네트워크를 이용하여 더욱 음질이 뛰어난 음성 변환을 한 연구도 존재한다 [11].

가장 최근에는 이미지 스타일 변환에서 사용하던 방법을 적용하여 인스턴스 정규화 방법으로 음성 변환을 직접적으로 한 연구도 존재한다. 이 연구 역시 언어적 정보와 스타일 정보를 나누기 위한 네트워크를 설계하는데, 기존 연구들은 스타일 정보를 제대로 추출해내지 못했던 반면에 이 연구는 스타일 정보를 어느정도 추출해낼 수 있었고 가능성을 제시하였다. 향후 음성 변환 연구들이 이런 방향으로 연구될 것으로 예상된다.

2.3 변분 자동인코더

2.3.1 목적 함수

변분 자동인코더는 잠재 변수 모델(Latent Variable Model)의 일종이다. 잠재 변수 모델이란 잠재 변수 z가 주어졌을 때의 조건부 확률 $P(X|z;\theta)$를 모델링하는 모델을 일컫는다. 실제로 우리가 원하는 데이터의 분포 $P(X)$를 다음과 같이 표현할 수 있다.

$$P(X) = \int P(X|z;\theta)P(z)dz \tag{2.2}$$

변분 자동 인코더에서는 가우시안 분포를 위의 $P(z)$로 사용한다. 즉 구하고자 하는 모델에서 $P(z)$를 알고, 신경망을 이용하여 $P(X|z)$를 모델링 한다면 실제 데이터 분포 $P(X)$를 구할 수 있다.

하지만 실제로 z에 대한 어떤 학습도 없이 가우시안 분포에서 무작위로 샘플링하여 사용하게 되면 $P(X|z)$의 값이 대부분 0이기 때문에, $P(X)$를 구하는 것이 무의미하다. 이를 위해서 샘플링된 z가 X를 잘 생성할 수 있도록 학습을 시켜주어야 한다. 즉, 완전히 무작위적인 z를 샘플링하는 것이 아니라, X를 잘 생성할 수 있는 z를 샘플링하겠다는 것이 목적이다. 이것은 확률 분포 $P(z|X)$로 표현할 수 있다. 이를 위해서는 변분 자동인코더의 인코더 부분에 대한 학습이 필요하다. 즉 변분 자동인코더의 전체적인 구조는 X를 잘 생성할 수 있는 z, 즉 $P(z|X)$를 학습하는 인코더와 $P(X|z)$를 학습하는 디코더의 구조로 나뉜다.

변분 자동 인코더를 학습하기 위한 목적함수는 ELBO [1]라 불리우며, 최대가능도 방법을 이용하여 학습한다. 모델이 배우고자 하는 식은 다음과 같다.

$$maximize(\log P(X) - D[Q(z|X||P(z|X)]) \tag{2.3}$$

즉, 주어진 데이터의 가능도 $P(X;\theta)$를 최대화하며 실제 $P(z|X)$와 인코더가 배우고자하는 분포 $Q(z|X)$의 거리를 가깝게 만들어줄 쿨벡-라이블러 발산값을 최소화시키는 것이 목적이다. 하지만 위 식은 실제 분포 $P(X;\theta)$를 모델링하는 것이 불가능하기 때문에 계산할 수 있는 형태로 변형을 거칠 필요가 있다.

$$\log P(X) - D[Q(z|X||P(z|X)] = \log P(X) - E_{z\sim Q}[\log Q(z|X) - \log P(z|X)] \quad (2.4)$$

베이즈 정리에 의해 $P(z|X)$는 다음과 같이 변형할 수 있다.

$$\log P(X) - E_{z\sim Q}[\log Q(z|X) - \log P(X|z) - \log P(z)] - \log P(X) \quad (2.5)$$

최종적인 목적함수는 다음과 같다.

$$E_{z\sim Q}[\log P(X|z)] - D[Q(z|X)||P(z)] \quad (2.6)$$

위의 식은 Q의 분포를 따르는 z에 대해 모델링이 가능하고, 따라서 계산이 가능하다.

2.3.2 최적화

목적함수를 최적화하기 위한 방법으로 경사하강법(Gradient Descent Method)를 사용한다. 모델의 구조 상 인코더에서 학습된 $Q(z|X)$의 분포에서 해당 z를 샘플링하는 과정에서 경사하강법의 흐름이 끊어지게 된다. 따라서 경사하강법의 흐름을 이어주기 위해서 재 매개변수화 속임수(reparameterization trick)라는 특별한 방법을 사용한다.

그림 2.2와 같이 학습된 $Q(z|X)$의 $N(\mu,\sigma)$의 분포에서 샘플링하는 것이 아니라, $\epsilon \sim N(0,1)$를 이용하여 $z = \mu + \sigma * \epsilon$으로 z를 샘플링하게되면 경사하강법의 흐름이 끊어지지않으면서 학습을 시킬 수 있게된다.

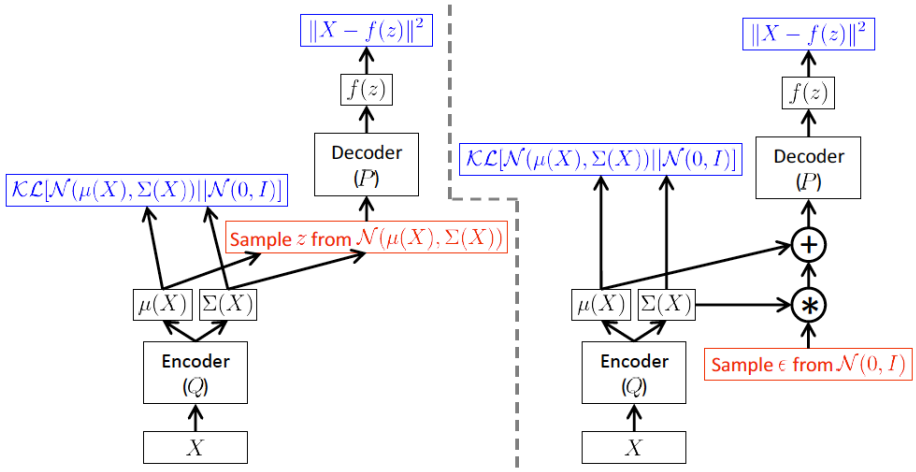

Figure 2.2: 재 매개변수화 속임수 [1]

2.3.3 조건부 변분 자동인코더

조건부 변분 자동인코더는 변분 자동인코더에 고정된 조건을 추가로 부여하기 위해 사용되는 방법이다. 목적함수는 거의 동일하며 단지 조건부 분포로써의 파라미터가 추가된다. 최종 식은 다음과 같다.

$$E_{z\sim Q}[\log P(Y|z, X)] - D[Q(z|Y, X)||P(z|X)] \tag{2.7}$$

여기서 X는 고정된 조건부 확률변수이며, Y는 실제로 생성하고자 하는 데이터의 분포이다.

2.4 역 자기회귀성 유동

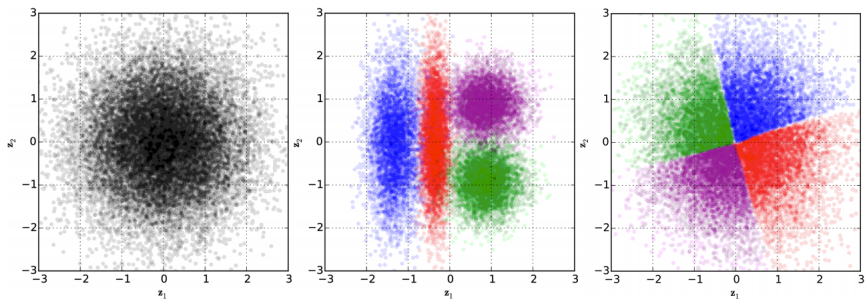

Figure 2.3: 역 자기회귀성 유동의 효과 [2]

확률적 변분 추론(Stochastic Variational Inference)는 연속적인 잠재 변수에 대해 효율적인 학습이 가능한 방법으로, 추론 네트워크를 학습한다는 것은 곧 사후 확률분포를 학습하는 과정이라 할 수 있다. 본 연구에서 사용하는 변분 자동 인코더는 변분 추론 학습 모델로써, 추론 네트워크와 생성 네트워크로 나뉘게된다. 기존의 변분 자동 인코더의 경우 추론 네트워크의 사후 확률 분포를 가우시안 분포로 가정하였다. 하지만 모든 사후확률 분포를 가우시안 분포로 가정하게되면 복잡한 사후확률 분포에 대해 제대로 추론하지 못하는 결과를 얻게 된다.

그림 2.3을 보면 가장 왼쪽의 그림이 사전 확률분포이며, 중간 그림이 가우시안으로 사후 확률분포를 가정할 경우에 대한 추론 결과이다. 이 경우에 사전 확률 분포와 유사하게 분포를 이루지 못할 뿐만 아니라 각각의 사후 확률분포도 가우시안의 형태로 남아있게되어 실제 분포와 달라지는 경향을 보인다. 이를 해결하기 위해 더욱 유연한 추론 네트워크를 만들고자 역 자기회귀성 유동을 사용한다. 역 자기회귀성 유동을 사용하게 되면 그림 2.3의 가장 오른쪽 그래프와 같이 사전 확률분포와 비슷하게 추론하면서도 사후 확률분포 역시 복잡한 분포로 추론이 가능하다. 관련 수식은 다음과 같다.

$$\log q(z_T|x) = \log q(z_0|x) - \sum_{t=1}^{T} \log \det \left| \frac{dz_t}{dz_{t-1}} \right| \tag{2.8}$$

위 수식에서 z_T는 역 자기회귀성 유동을 거친 사후확률분포를 나타내며 z_0는 기존의 가우시안 확률 분포이다. 이 아이디어는 [15]에서 사용하였지만, 일반 비선형 함수의 경우 위 수식의 마지막 부분인 자코비안 행렬의 결정자값을 구하는 계산에 드는 비용이 매우 높았기 때문에 일부 가능한 함수들만 위 수식으로 분포를 변형할 수 있었다. 하지만 역 자기회귀성 유동을 사용하게 되면 선형 변환 후 비선형 함수로 활성화하는 평범한 네트워크 구조인 함수에 대해 자코비안 행렬의 결정자값을 쉽게 구할 수 있기 때문에 분포의 변형도 간단하다. 즉, 한 번의 역 자기회귀성 유동이 신경망으로 대체될 수 있다. 추론 모델에 대해 역 자기회귀성 유동을 적용한 최종 식은 다음과 같다.

$$z_{t-1} = (z_t - \mu(z_t))/\sigma(z_t) \tag{2.9}$$

$$\log \det \left| \frac{dz_{t-1}}{dz_t} \right| = -\sum_{i=1}^{D} \log \sigma_i \tag{2.10}$$

$$\log q(z_T|x) = -\sum_{i=1}^{D} (\frac{1}{2}{\epsilon_i}^2 + \frac{1}{2}\log(2\pi) + \sum_{t=0}^{T} \log \sigma_{t,i}) \tag{2.11}$$

여기서 D는 자코비안 행렬의 차원을 뜻하며, $\sigma_{t,i}$는 t번째 역 자기회귀성 유동 신경망의 i번째 차원에 대한 값을 뜻한다.

[2]에서는 변분 자동 인코더의 학습 개선을 위해 프리빗 알고리즘이라는 새로운 알고리즘을 소개하고 있는데, 이 방법은 사전 확률분포와 사후 확률분포가 초기에 너무 빨리 가까워져 쿨벡-라이블러 발산값이 0에 너무 빨리 수렴하는 현상을 막기 위함이다. 초기에 쿨벡-라이블러 발산이 0에 수렴하게 되면 모델이 사후 확률분포에 대한 정보를 학습하지 못하고 사전 확률분포에 머무르게 되는 현상이 발생하게 되는데, 이를 방지

하기 위함이다. 이 알고리즘의 핵심은 쿨벡-라이블러 발산의 하한값을 임의의 값으로 고정한 후 학습하는 것이다. 프리빗 알고리즘에 의해 변형된 쿨벡-라이블러 발산값은 다음과 같이 계산된다.

$$D'_{KL} = \sum_{j=1}^{K} maximum(\lambda, E_{x\sim M}[D_{KL}(q(z_j|x)|p(z_j))]) \tag{2.12}$$

위 식에서 D_{KL}은 기존의 쿨벡-라이블러 발산을 나타내며, K개의 소그룹으로 잠재 변수의 차원을 나눈 뒤에, 각각의 값이 임의의 λ 이상일 경우만 그 값을 취하는 형식이다.

2.5 관련 연구

[8]는 두개의 모듈로 구성된 심층 인공 신경망을 사용하는 음소 기반 다대일(many-to-one) 음성 변환 시스템을 제안했다. 첫번째 모듈은 원본 화자의 발화에서 음소를 인식하는 것이고, 두번째 모듈은 인식된 음소를 기반으로하여 목표 화자의 발화를 생성하는 것이다. 그러나 이 모델의 경우 변환 단계에서 원본 화자의 발화에 따라 단 하나의 발화만을 생성 할 수있는 결정론적 모델이다. [16]는 원본 화자와 목표 화자의 발화에 대해 동일 음소의 발화 시간이 일치해야한다는 제한을 극복하기 위해 시퀀스-시퀀스 모델(Sequence-to-Sequence Model)을 도입하여 모델을 개정하였다.

일부 연구에서는 변분 자동인코더 및 생성형 적대 신경망과 같은 생성 모델을 사용하여 비감독형 음성 변환 방법을 제안하였다. [10]는 변분 자동인코더를 이용하여 오디오에서 얻은 잠재변수에 대해 대상 화자의 화자 식별자를 조건으로 추가하여 음성을 변환하는 방법을 제안하였다. 즉, 이 모델에서는 오디오의 언어적인 부분을 잠재변수에 대응하여 학습하도록 하였다. [11]는 [10]에서는 보다 현실적인 음성을 생성하기 위해 Wasserstein-GAN을 사용하여 성능을 더욱 향상시켰다. [17]에서 그들은 변분 자동인코더의 잠재변수에 벡터 양자화 기법을 사용하여 유한 수의 잠재 성을 배워 음성으로부터 내용을 추출하는 방법을 제안했다. 위의 모든 연구는 생성 모델을 사용하여 비감독 방식으로 언어적 정보를 학습하는 방법을 제시하지만 이 모델의 목표인 목표 화자의 다양한 억양을 갖는 발화를 만들 수 없다는 점에서 한계가 존재한다.

[4], [18], [19], [12], [20] 및 [21]와 같은 텍스트를 기반으로 음성을 생성하는 음성 합성 시스템에서 많은 연구가 수행되었다. 텍스트에서 직접적으로 오디오를 생성하는 [22]도 소개되었다. [19]에서는 매 시간 단계마다 억양에 해당하는 정보와 기본 주파수에 해당하는 음소 정보를 조건으로 하여 음성을 생성하였다. 그러나 본 연구는 매 시간 단계마다 다른 기본 주파수값이 아닌 모든 시간 단계에 걸쳐 억양 정보를 담고있는 단

하나의 잠재적인 벡터로 억양을 다양화하는 방법을 제안하였다.

제 3 장 제안 기법

제안된 방법은 다대일 모델을 기반으로하며, 이는 모든 화자의 발화를 단일 화자의 발화로 변환하는 모델이다. 모델은 크게 음소 분류기와 음성 합성기로 구조화된다. 첫째로, 음소 분류기는 원천 화자의 발화로부터 화자의 발화 스타일과는 독립적인 언어 특성을 추출하여 목표 화자의 발화를 생성하기 위한 조건으로 사용한다. 음성 합성기는 추가적인 확률론적 변수를 입력으로 받아서 다양한 음성 스타일을 생성한다. 음소 분류기는 음성 합성기가 학습되기 이전에 미리 학습시켜 놓는 방식으로, 각 모듈을 개별적으로 학습한다. 전체적으로, 음소 분류기에서 음성 프레임을 입력으로 받아 각 프레임에 대한 음소를 예측하고, 이 예측된 음소를 조건으로 하여 조건부 변분 자동인코더 구조를 가진 음성 합성기를 통해 최종적으로 변환된 음성을 생성한다. 음성 합성기는 목표화자의 목소리로만 학습시키게 되며, 음소 분류기는 다양한 화자의 목소리에 대한 데이터를 바탕으로 학습시킨다.

3.1 음소 분류기

음소분류기는 각 프레임을 음소의 확률값으로 변환한다. 분류기의 마지막 소프트맥스 레이어 직전의 로짓은 언어적인 정보만을 담고있는 것으로 가정한다. 주어진 음성 스펙트로그램 프레임들과 원-핫 인코딩(ont-hot encoding)된 데이터 쌍은 $x_{s,t}, y_{s,t}$로 표현된다. 이때 아랫첨자 s는 화자를 나타내며, t는 시간 단계를 나타낸다. 음소분류기는 언어적 특징을 다음과 같이 추출한다.

$$c_{s,t} = f(x_{s,t}, h_{t-1}) \tag{3.1}$$

여기서 f는 음소 분류기를 나타내며, h는 은닉 상태를, c는 각 시간 단계 t의 프레임에 대한 확률 벡터를 나타낸다. 학습 단계에서, 가능도를 최대화하기 위해 크로스-엔트로피 목적함수를 사용한다. 관련 수식은 다음과 같다.

$$L_{PC} = \sum_{t=1}^{T} CELoss(c_{s,t}, y_{s,t}) = \sum_{t=1}^{T} -y_{s,t} \log c_{s,t} \tag{3.2}$$

L_{PC}는 음소 분류기의 손실 함수를 뜻한다.

3.2 음성 합성기

음성 합성기는 음소 분류기로부터 출력된 음소의 확률을 입력으로 취하며, 매 타임 스텝마다 목표 화자의 발화 스펙트럼을 추정한다. [8]에서 영감을받은 결정론적 기준 모델을 베이스라인 모델로 삼아 제안 모델의 효과와 비교한다. 베이스 라인 모델은 단순히 [4]이 제안한 두 개의 CBHG 모듈로 구성된다. 음소 분류기에서 추출한 언어적 특징을 조건으로 사용하는 조건부 변분 자동인코더 (CVAE) 기반 모델을 제안한다. 조건부 변분 자동인코더는 인코더인 추론 모델과 디코더인 생성 모델로 구성된다. 추론 모델은 목표 화자의 모든 시간 단계에 걸친 음소 확률 벡터의 연결 벡터인 c_{tgt}와 음성 프레임 x_{tgt}가 입력으로 주어지고, 사후확률 z_{tgt}을 추정하는 모델이다. 관련 수식은 다음과 같다.

$$\mu_{tgt}, \sigma_{tgt} = g(x_{tgt}, c_{tgt}) \tag{3.3}$$

$$z_{tgt} = \mu_{tgt} + \sigma_{tgt} * \epsilon \tag{3.4}$$

여기서 g는 조건부 변분 자동인코더의 추론 모델을 나타내며, ϵ 은 평균 0, 분산 1인 표준 정규분포에서 샘플링된 값이다. 최종적으로, 이미 고정된 언어적 정보를 담고있는 c_{tgt}과 사후확률인 z_{tgt}을 이용하여 생성모델 h가 다양한 억양을 가지는 스펙트로그램들을 생성한다.

$$\hat{x}_{tgt} = h(z_{tgt}, c_{tgt}) \tag{3.5}$$

조건부 변분 자동인코더의 학습 단계에서 가능도를 직접적으로 계산하는 것이 불가능하기 때문에, ELBO라 불리는 가능도의 최소 경계값을 이용하여 이 값을 최대화

함으로써 모델을 학습한다. ELBO를 구하는 식은 다음과 같다.

$$ELBO = \mathbb{E}_{z\sim Q(\cdot|C,X)}[\log P(X|z,C)] - D[Q(z|C,X)||P(z|C)] \tag{3.6}$$

여기서 X는 각 음성 프레임에 대한 확률 변수를 의미하며, C는 음소 확률에 대한 확률 변수를 의미한다. Q는 사후 확률의 근사값을 의미하며 D는 쿨벡-라이블러 발산을 의미한다. 이때, z가 C에 대해 독립적으로 샘플링되기 때문에, 본 모델에서는 사전 확률로 $P(z|C)$ 대신 $P(z)$를 사용하였다. ELBO에 따라 재구성 손실함수 L_{rec}와 쿨벡-라이블러 발산 값 L_{kld}을 포함한 음성 합성기의 전체 손실함수를 계산하면 다음과 같다.

$$L_{rec} = ||x_{tgt} - \hat{x}_{tgt}||_2^2 \tag{3.7}$$

$$L_{kld} = \frac{1}{2}(\mu_{tgt}^2 + \sigma_{tgt}^2 - 2\log\sigma_{tgt} - 1) \tag{3.8}$$

$$L_{SS} = L_{rec} + L_{kld} \tag{3.9}$$

여기서 σ_tgt는 추론 모델을 통과하여 출력된 잠재 변수의 분산 값이다.

3.3 샘플 생성

추론시, 원천 화자, 즉 변환 대상이 되는 발화는 먼저 스펙트로그램으로 변환되어 음소 분류기에 입력값으로 넣어주며, 출력값으로 언어적 정보를 담는 벡터를 얻는다. 그런 다음, 분포 $N(0, I)$에서 잡음 벡터를 샘플링한다. 잡음 벡터 및 분류기로부터 추출된 언어적 정보를 담은 벡터는 입력값으로써 음성 합성기를 통과한다. 다양한 억양의 변환된 발화를 얻기 위해 무작위로 잡음을 반복적으로 샘플링한다. 이렇게 생성된 스펙트로그램을 원시 파형으로 복원하기 위해 그리핀-림 알고리즘 (Griffin-Lim Algorithm) [14]. 이때 추가적으로 생성되는 잡음을 제거하기 위해 예측된 스펙트로그램에 1.2의 거듭제곱을 취한다.

3.4 역 자기회귀성 유동 적용

보다 다양한 억양을 가지는 음성을 생성하기 위해 기존의 변분 자동 인코더에서 적용한 표준 정규분포에서 ϵ을 샘플링하는 것 보다 더 복잡하고 유연한 사후 분포를 만드는 것의 필요성이 있다. 역 자기회귀성 유동은 초기 표준 정규분포에서 반복적이고 가역적인 변환 체인을 사용하여 더욱 복잡한 사후 확률 분포를 만드는 일종의 정규화 흐름이다. 정규화 흐름에 따르면 확률밀도 함수는 다음과 같이 전개된다.

$$\log q(z_T|x) = \log q(z_0|x) - \sum_{t=1}^{T} \log \det \left| \frac{dz_t}{dz_{t-1}} \right| \tag{3.10}$$

여기서 z_0는 3.2절에서 설명한 z_{tgt}, 즉 초기 잠재 벡터이며 z_T는 역 자기회귀성 유동을 적용한 최종 벡터이다. 위 방정식을 계산하기 위해서는 반복 횟수 T에 대해 자코비안 행렬식을 계산하는 것이 필수적이다. 그러나 자코비안 행렬식을 계산하는 것은 계산 상 매우 비싸다. 따라서 하방삼각 가중치 행렬을 갖는 자동 회귀 네트워크가 결정자의 계산을 단순화 하기 위해 구성되었다.

$$z_{t-1} = (z_t - \mu(z_t))/\sigma(z_t) \tag{3.11}$$

$$\log \det \left| \frac{dz_{t-1}}{dz_t} \right| = -\sum_{i=1}^{D} \log \sigma_i \tag{3.12}$$

이 방법을 이용하면, 각 시간 단계에서 z, μ, σ 사이의 관계를 이용하여 사후 확률 밀도 함수를 쉽게 계산할 수 있다. 역 자기회귀성 유동을 이용한 본 모델의 음성 합성기에 대한 최종 손실함수는 다음과 같이 계산된다.

$$L_{SS} = ||x_{tgt} - \hat{x}_{tgt}||_2^2 + \frac{1}{2}(z_T^2 - \epsilon^2 - 2\sum_{t=1}^{T} \log \sigma_t) \tag{3.13}$$

여기서 z_T는 역 자기회귀성 유동을 통과한 최종 잠재 벡터이며, σ_t는 각 t번째 역 자기회귀성 유동 네트워크의 분산 값이다.

3.5 모델 구조

음소 분류기의 구조는 128개의 은닉 차원, 8개의 1차원 합성곱 필터 뱅크, 4개의 하이웨이 블록 (highway block) [23] 및 드롭 아웃 비율 0.2를 포함하는 CBHG 모듈로 구성된다. CBHG는 기본적으로 다양한 커널 크기의 합성곱 계층을 사용하며 gated recurrent unit (GRU)이 추가된 메모리 셀이다.

2개의 양방향 LSTM 층과 256 개의 은닉 노드로 구성된 2개의 완전히 연결된 (fully-connected) 은닉층으로 구성된 추론 모델이다. 그런 다음 마지막으로 완전히 연결된 은닉층의 출력은 z 의 평균 μ과 분산 σ을 나타내는 두 개의 16 차원 벡터로 투영된다.

생성 모델은 256개의 은닉 차원, 8개의 1차원 합성곱 필터 뱅크, 8개의 하이웨이 블록 및 드롭 아웃 비율 0.2를 포함하는 두 개의 CBHG 모듈로 구성된다. 출력은 선형 스펙트로그램에 투사된다. 역 자기회귀성 유동을 사용하는 모델의 경우 12 개의 역 자동 회귀 선형 은닉층을 사용하였다. 조건부 변분 자동인코더를 훈련하는 동안 쿨벡-라이블러 발산이 너무 빨리 수렴되는 것을 제한하기 위해 [2]에 언급 된 프리빗(freebits) 알고리즘을 사용하였다. 세부 하이퍼 파라미터는 $\lambda \in [0.125, 0.5, 1.0, 2.0]$과 $K \in [4, 8, 16]$을 사용했다. 모델의 전체적인 구조는 그림 3.1에 묘사되어있다.

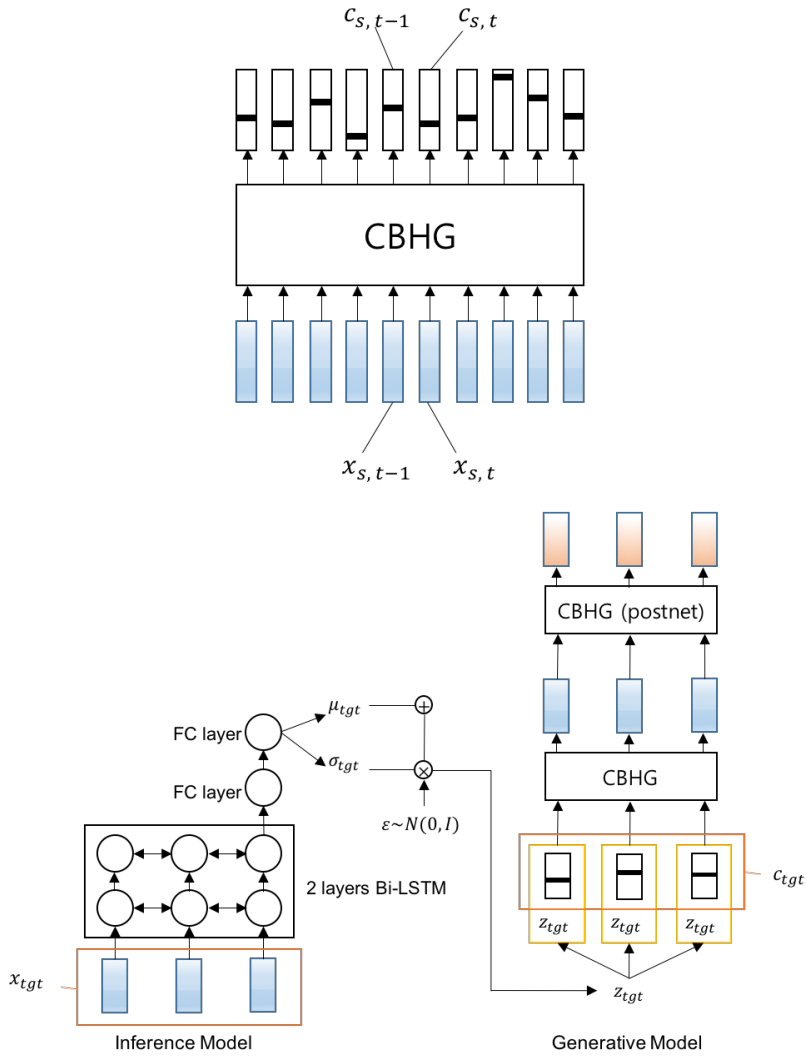

Figure 3.1: 전체 모델 구조

제 4 장 실험 및 결과

4.1 실험 설계

음소 분류기는 TIMIT 코퍼스 [24]에서 음소적으로 균형 잡힌 6,300 개의 발화와 미국 영어의 8개 주요 방언의 630개 발화자 데이터셋에 대해 학습되었다. 이 데이터 셋은 각 음성 파일에 대한 문자열 스크립트와 함께 음소쌍이 제공된다. 고유 음소의 갯수는 총 61개로 구성되어 있다. 이 데이터는 각 음소가 음성 데이터에서 어느 시간까지 발화되고 있는지에 대한 정보를 담고있다. 예를 들어, 0s부터 100ms까지는 's', 100ms부터 500ms까지는 'ah'가 발화된다는 정보를 담고있다. 이 데이터셋의 샘플링률은 16,000Hz 이며 주로 음성 인식분야에서 널리 쓰이는 데이터셋이다.

조건부 변분 자동인코더 기반 음성 합성기는 Librispeech 데이터셋 [25] 중 하나인 LJ Speech 데이터셋을 이용하여 학습하였다. LibriSpeech 데이터셋 역시 음성 인식 분야에 널리 쓰이는 데이터셋으로, 총 약 1,000시간의 발화와 이에 대한 스크립트로 구성되어 있다. 샘플링률은 16,000Hz이며 잡음이 없는 깨끗한 데이터셋과 잡음이 포함된 데이터셋으로 구성되어 있다. 모든 음성 데이터는 영어로 발화되었다. LJ Speech 데이터셋은 7 권의 논픽션 책자를 읽는 단 한 명의 화자에 대해 13,100개의 짧은 오디오 클립으로 구성되어있으며 총 길이는 약 24 시간이다. 경험적으로, 단일 화자의 목소리에 대해 학습하기 위해 최소 12 시간의 데이터가 필요하고 본 연구의 목표화자가 단 한명이기 때문에 LJ Speech 데이터셋을 사용하였다.

본 논문에서는 CMU Arctic 코퍼스 [26]를 원천 발화로 사용하였다. 이 데이터셋은 약 1,200 개의 음성학적으로 균형 잡힌 영어 발화가있는 18 개의 단음 연설 데이터베

Table 4.1: 음성데이터 전처리 하이퍼-파라미터

종류	값
샘플링 률	16000(Hz)
분할 시간	2(sec)
윈도우 크기	25(ms)
홉 크기	5(ms)
멜 필터뱅크 채널 수	80(개)
MFCC 개수	40(개)
사전 증폭 비율	0.97
스팩트로그램 채널 수	256
정규화 최대 데시벨	35(db)
정규화 최소 데시벨	−55(db)

이스로 구성되어있다. LJ Speech 데이터가 여성의 목소리이기 때문에 이중에서 여성 화자와 남성 화자에 대한 데이터를 모두 사용하여 여성-남성간의 음성 변환, 여성-여성간의 음성 변환에 대해 실험하였다.

음성 데이터의 전처리 과정은 표 4.1에 정리되어있으며 세부사항은 다음과 같다. 우선 샘플링 비율 16kHz로 샘플링된 파형을 2초 단위로 분할하고, 계수 0.97로 사전 증폭하였다. 멜 스펙트로그램은 25ms 프레임 크기, 5ms 프레임 홉 및 한(Hann) 윈도우 함수를 사용하여 단시간 푸리에 변환 (STFT)을 통해 계산되었다. 그 다음, 80 채널을 가지는 멜 필터 뱅크를 사용하여 STFT 크기를 멜 스케일로 변환하였다. 이후 셉스트랄 분석을 통해 멜 스케일된 스펙트로그램에서 변형된 40 개의 멜 주파수 셉스트랄 계수 (MFCC)를 음소 분류기의 입력값으로 사용하였다.

훈련 과정에서 사용된 최적화 도구는 다음과 같다. 우선 음소 분류기의 경우 최적화를 위한 경사하강법 최적화 도구로써 Adam optimizer [27]을 $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$ 및 3e-4의 고정된 학습률로 사용하였다. 훈련된 음소 분류기로부터 추론된 해당 음소 레이블을 조건부 변분 자동인코더 기반 음성 합성기에 대해 학습시켰다. 음소 분류기의 학습 시와 동일한 Adam 최적화 도구를 사용하여 배치 크기를 32로 사용하였

다. 학습률의 경우 조정을 거쳤는데, 초기 학습률을 5e-4로 고정시킨 다음 150,000 회 반복한 후 1e-4로 조정하였다. 모든 실험 및 코드는 Python 및 Pytorch로 구현되었다.

4.2 음소 분류기 학습 결과

음소 분류기 학습 결과 가장 성능이 좋은 분류기는 73 퍼센트의 정확도를 보였다. 그림 4.1은 예측한 음소들의 확률값을 각 프레임별로 도식화 한 것의 예시이다. 가로축은 각 프레임들의 시간축을 나타내며, 세로축은 61개의 음소들을 나타낸다. 각 픽셀들은 0 1사이의 확률값들을 가지며 세로축으로 모든 픽셀을 더할 경우 합이 1이 된다. 픽셀의 색이 하얀색에 가까울수록 확률값이 1에 가까움을 의미한다. 위 혼동행렬에서 볼 수 있듯이 확률값들이 여러 값으로 산포되어있는 것이 아니라 특성 음소에 집중적으로 높은 것을 확인할 수 있다.

Figure 4.1: PPGs 샘플

혼동 행렬을 기반으로 분석을 진행한 결과, 분류기의 가장 빈번한 오답의 경우 ’s’/ ’z’, ’ah’/ ’aa’, ’ae’/ ’eh’및 ’ax’/ ’ix’인 것을 발견하였다. 이 음소쌍들은 거의 같은 발음을 가리키며, 인간이 판단하기에도 맞추기 어려운 음소쌍들이다.

음소 분류기의 학습 성능을 더 높이기위한 시도는 하지 않았는데, 이는 본 논문의 목적이 음소 분류기의 학습 성능을 높이는 것이 아니라 같은 음소 확률 벡터에서 다양한

억양을 가지는 발화를 생성해낼 수 있는가이기 때문이다. 데이터 양을 늘리거나 모델의 구조를 바꾸는 등 음소 분류기의 성능을 높일 수 있는 여지가 남아있지만 이는 향후 연구로써 남겨두었다. 모델 구조를 더 복잡하거나 자기회귀성 구조를 가지도록 하게되면 음소의 특성상 앞에 어떤 음소가 나왔느냐가 다음 음소 예측에 매우 좋은 정보로 작용할 것이기 때문에 성능 향상이 예상된다.

4.3 다양한 억양

본 논문에서 제안된 모델이 멜 스펙트로 그램을 플로팅하고 변환된 샘플을 직접 듣는 방식을 통해 다양한 억양을 가진 발화를 생성함을 확인했다. 결과는 ϵ이 μ에서 보다 더 멀리 있는 값에서 샘플링될수록 모델이 훨씬 더 다른 억양으로 발화를 생성함을 보여주었다. 여기서 억양의 변화란, 똑같은 문장을 발음하지만 발화 중간의 단어간 또는 단어 내의 음의 높낮이가 달라짐을 의미한다. 그러나 ϵ 가 3σ 이상의 분포에서 샘플링되면 언어 정보가 손실되어 합성된 발화가 올바른 문장을 나타내지 않는 것을 확인하였다. 이는 μ값에 가까울수록 가장 일반적인 해당 화자의 발음, 억양에 대한 정보를 바탕으로 생성하게 되고, 2 3σ사이에 있는 샘플의 경우 학습할 때 입력값으로 충분히 접하지 못했지만 해당 화자가 간헐적으로 사용하는 억양에 대한 정보를 학습한 것으로 사료된다. 그리고 3σ 이상의 샘플의 경우 학습시 거의 이 데이터를 접해보지 못했기 때문에 완전히 새로운 분포를 생성하여 언어적인 특성에도 악영향을 끼치는 것으로 판단된다.

또다른 실험으로, ϵ의 변화로 인해 억양 변화가 어떤 방식으로 변하는 지 보간 실험을 수행하였다. 먼저, 가장 다른 억양을 가진 두 개의 발화를 샘플링하여 ϵ 값을 측정한 다음, ϵ을 다음과 같이 변경하면서 실험을 수행하였다.

$$\epsilon = \alpha\epsilon_1 + (1-\alpha)\epsilon_2 \tag{4.1}$$

여기서 α는 [0,1]의 범위 내에 존재하는 실수값이다. 결과적으로 멜 스펙트로그램을 ϵ을 변화시켜가며 그려보았을 때, 그 변화에 따라 연속적으로 변화하는 것을 확인할 수 있었다. 변화를 보여주는 그림은 그림 4.2, 그림 4.3에서 확인할 수 있다. 그림을 보면, 전반적인 형태는 비슷하지만 시간에 따른 변화가 올라가거나 내려가느냐에 따른

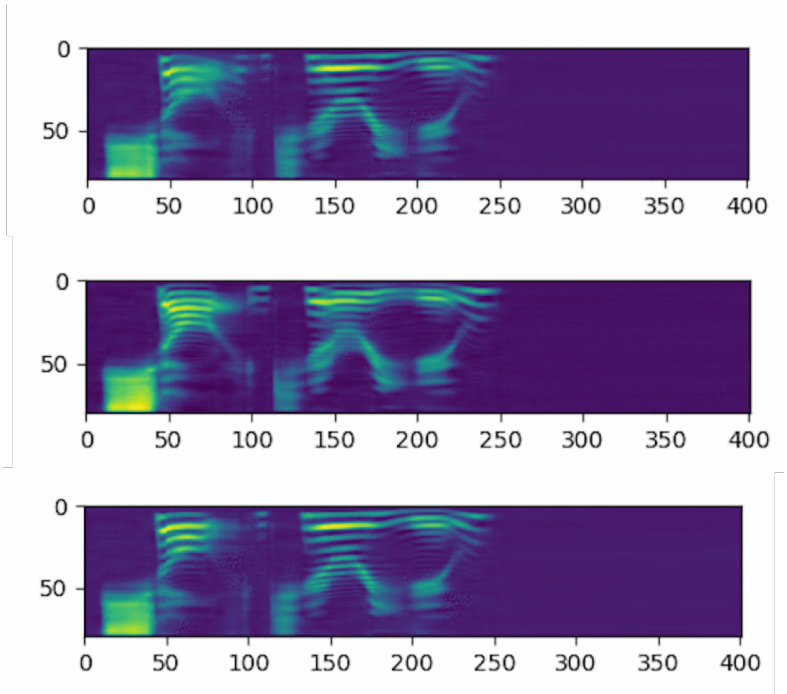

Figure 4.2: ϵ에 따른 억양의 변화

추세가 달라짐을 확인할 수 있고, 이것이 억양의 변화로 이어짐을 확인하였다.

그러나 샘플을 직접 들었을 때, ϵ의 변화로 인한 억양의 명백한 변화를 인식하는 것은 어려웠다. α를 10개로 나누어 샘플을 들어봤을 때, 억양의 변화가 특정한 경계 근처에서 이산적으로 일어나는 것을 확인할 수 있었다. 이는 사람이 들을 수 있는 가청 범위 내에서는 이산적으로 변화하지만, 멜 스펙트로그램으로 확인해봤을 때 그 변화가 충분히 존재하는 것임을 알 수 있다. 멜 스펙트로그램의 gif 파일 및 샘플들의 변화 과정은 "https://soobinseo.github.io"에서 확인할 수 있다.

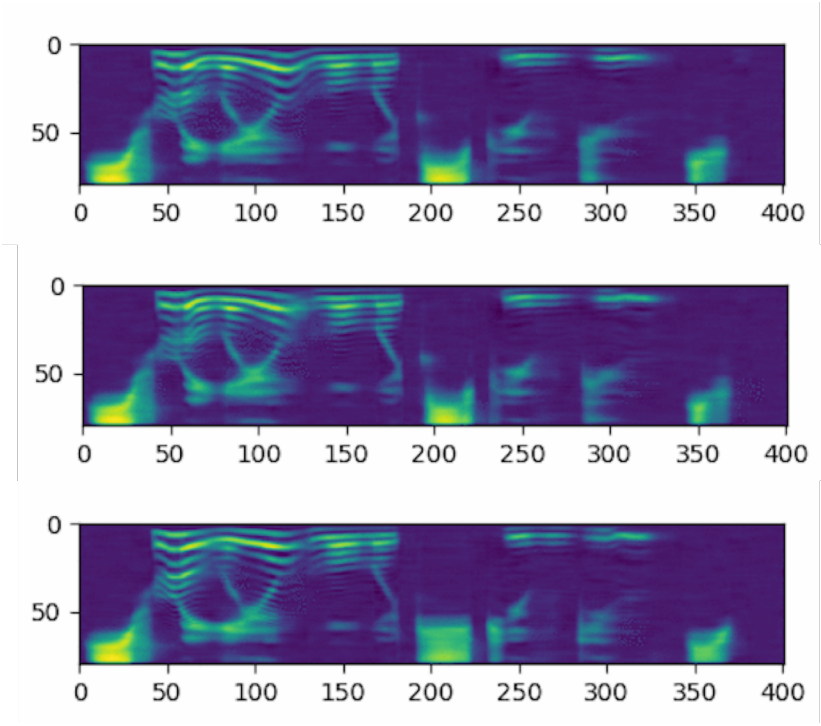

Figure 4.3: ϵ에 따른 억양의 변화 2

4.4 품질 평균 의견 점수 (MOS quality)

기본 모델과 비교할 때, 제안된 모델의 결과는 다양한 억양뿐만 아니라 더 나은 품질을 가지는 것을 확인하였다. 5-scale Mean Opinion Score (MOS)는 생성된 발화의 자연스러움을 평가하는 데 사용되었다. 기본 모델을 포함하여 바닐라 변분 자동 인코더와 역 자기회귀성 유동을 거친 변분 자동 인코더, 총 3가지 모델을 사용하여 실험하였다. 각 원천 발화는 CMU-arctic corpus에서 무작위로 뽑은 남성과 여성의 목소리이며, 남성과 여성 각각 2 개의 샘플을 음성 변환하였다. 이 세 가지 모델에서 생성 된 총 12개의 발화문과 8개의 정답 데이터셋이 무작위 순서로 제시되고 평가되었다. 정답 데이터셋은 원천 화자의 기존 데이터, 목표 화자의 기존 데이터에서 샘플링하여 제시되었다. 표 4.1에서 볼 수 있듯이, 변분 자동인코더를 사용하는 두 모델 모두 기준 모델보다 높은 점수를 보였다. 이는 변분 자동 인코더의 역할이 억양만을 변화시키는 것이 아니라 다양한 확률적 모델링을 통해 가장 자연스러운 스타일의 억양을 찾아서 음성을 생성하는 것으로 사료된다. 이 결과값들은 1σ 이하의 값에서 샘플링을 한 경우이며, 앞에서도 언급했듯이 μ에 가까울수록 해당 화자가 발화하는 모든 데이터들 중 자연스러운 경향을 학습한 것으로 사료된다.

Table 4.2: Mean Opinion Score (MOS)

Model	**MOS score**
Ground-truth (LJ Speech)	4.83 ± 0.07
Ground-truth (Arctic)	4.24 ± 0.18
Vanilla VAE	2.70 ± 0.25
VAE with IAF	2.47 ± 0.29
Non-VAE (baseline)	2.23 ± 0.27

다음 그림은 훈련이 진행되면서 멜-스펙트로그램의 변화를 나타낸다. 그림 4.4, 4.5, 4.6는 각각 10,000번, 30,000번, 50000번 훈련을 진행한 결과이며, 훈련을 거듭할수록

아래의 원본과 가까워짐을 확인할 수 있다. 약 100,000번 이상 훈련을 진행한 경우에는 더이상 음질의 향상이 없음을 확인하였다.

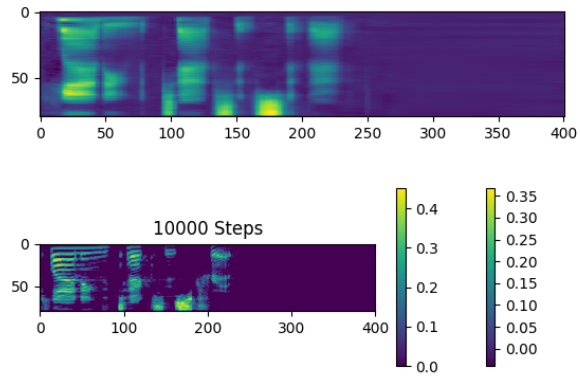

Figure 4.4: 10000번 훈련 결과

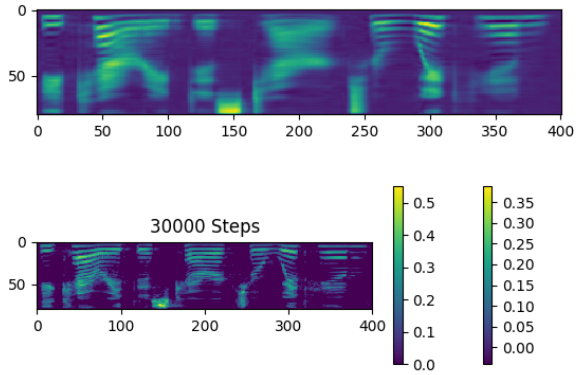

Figure 4.5: 30000번 훈련 결과

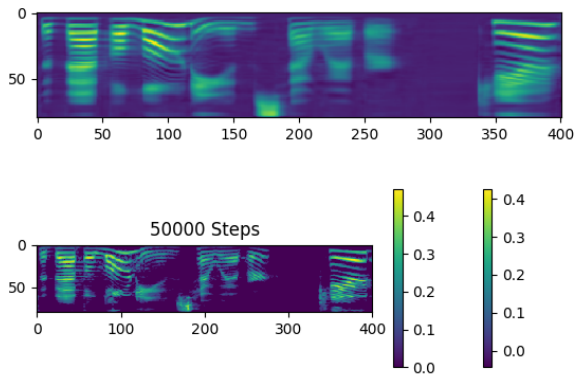

Figure 4.6: 50000번 훈련 결과

4.5 절제 연구

4.5.1 역 자기회귀성 유동의 효과

멜 스펙트로그램을 도식화하면서 동시에 샘플을 들어본 결과 멜 스펙트로그램이 활성화되는 구간은 비슷하지만 추세가 내려가는 추세이거나 올라가는 추세이냐에 따라 억양이 달라짐을 확인하였다. 그림 4.7에서 볼 수 있듯이, 전체 발화의 일부분만이 변분 자동인코더만을 사용할 때 z을 따라 약간 변화하지만, 역 자기회귀성 유동을 거친 z의 경우 발화가 시간축 전체적으로 변화하는 경향이 있음을 확인하였다. 역 자기회귀성 유동을 거치지 않은 Non-IAF의 그림을 보게되면 250 에서 300 시간축에 대해서만 내려가는 추세, 일직선인 추세임을 확인할 수 있다. 하지만 역 자기회귀성 유동을 적용한 IAF의 그림을 보면 시간축에 대해 전체적으로 추세가 변화한다. 또한 변환된 여러 개의 발음을 샘플링한 결과 역 자기회귀성 유동을 사용할 때 샘플 간의 억양의 차이를 증가시키는 것도 확인하였다. 즉, 추세의 기울기가 급격히 변한다거나 완전히 달라짐을 확인할 수 있었다.

이것은 z 가 역 자기회귀성 유동을 통화함으로써 근사한 사후분포가 더 복잡해짐에 따라 전체적으로 스타일에 대한 분포를 잘 근사하였고, 이것이 억양의 전체적인 변화에 영향을 미치는 것으로 분석된다.

4.5.2 멜 스펙트로그램과 선형 스펙트로그램의 차이

[4]에서는 멜 스펙트로그램과 선형 스펙트로그램을 주요 특징으로 사용했으며 멜 스펙트로그램을 사용하는 모델이 약간 더 나은 성능을 보였다. 데이터의 멜 스펙트로그램과 예측한 멜 스펙트로그램의 차이값을 손실함수로 사용한 이유는 멜 스펙트로그램이 사람의 가청 주파수대에 있는 특징들에 대한 정보를 많이 담고 있으며, 이렇게 예측된 멜 스펙트로그램을 바탕으로 음성을 생성한다면 사람이 듣기에 선명한 음성이 생성될

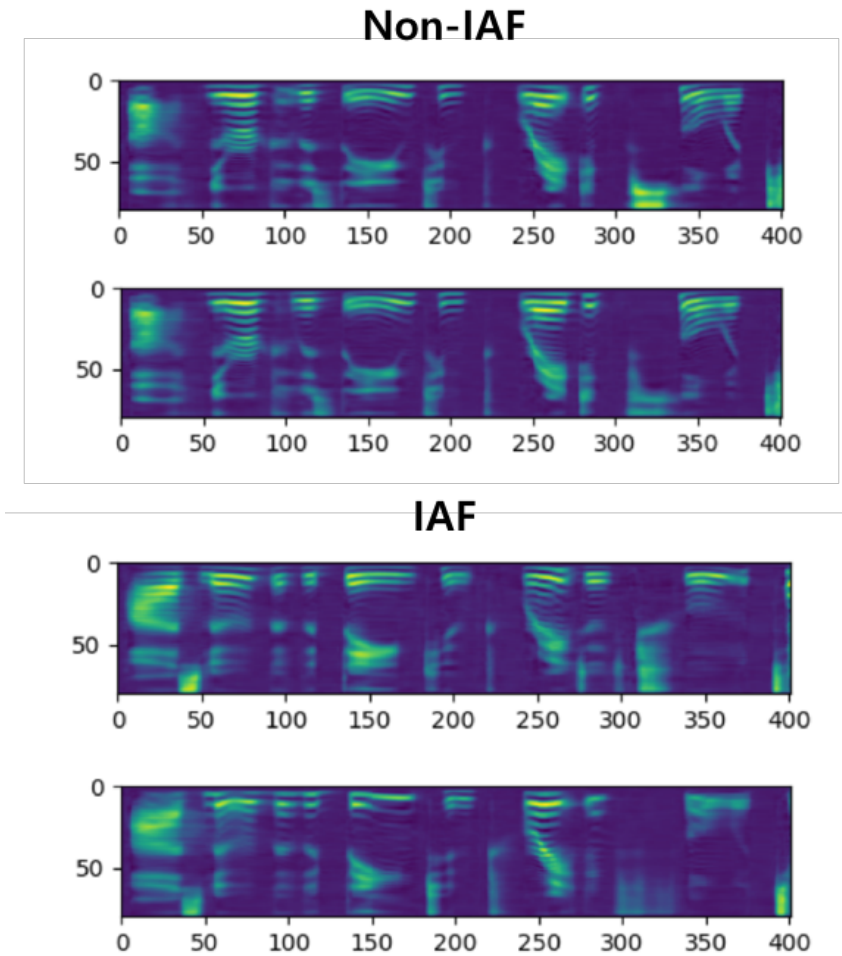

Figure 4.7: 역 자기회귀성 유동 적용 비교 실험 결과

것이라는 가정이 있기 때문이다.

본 논문에서 역시 동일한 절제 시험이 수행되었다. 첫 번째 모델은 멜 스펙트로그램을 사용하여 음성 합성기를 학습 한 후, 포스트 네트워크를 추가하여 선형 스펙트로그램을 출력하는 2 단계 모델이다. 대조적으로, 두 번째 모델은 음성 합성기를 교육하기 위해 선형 스펙트로그램을 직접 사용하였다. 그 결과 멜 스펙트로그램을 생성한 후 포스트 넷을 통해 선형 스펙트로그램을 생성하는 경우가 선형 스펙트로그램을 직접 생성하는 경우보다 음질이 좋아졌다. 이 결과는 약 10개의 샘플을 각 모델에서 추출하여 직접 청음한 결과이다. 대조적으로, 억양의 변화는 크게 다르지 않았다. 이것은 모델의 일반화 능력이 멜 스펙트로그램 사이의 재구성 손실함수로 인한 정규화 효과에 의해 증가되었다고 믿어진다.

4.5.3 프리빗 알고리즘

학습 시 사용되는 프리빗 알고리즘은 다양한 억양으로 음성을 변환하는데에 중요한 영향을 미쳤다. 프리빗 알고리즘은 사전 확률분포와 사후 확률분포간의 쿨벡-라이블러 발산에 대한 수렴 속도를 조절해주는 역할을 한다. 억양은 $\lambda < 1$에서 학습 할 때 거의 변하지 않았고 $\lambda > 1$에서는 ϵ에 따라 억양이 변화되었음을 확인할 수 있었다. 그림 4.8과 4.9을 보면 그 차이를 알 수 있다. 그림 4.8의 $\lambda = 1$의 경우 ϵ의 변화에 따라 멜 스펙트로그램의 변화가 아주 미세하거나 거의 없는 것을 확인할 수 있다. 이에 비해 그림 4.9의 경우 $\lambda = 2$일 때 시간축 150 200 사이에서 변화가 확연하게 보임을 알 수 있다. 이 샘플을 직접 들어보았을 때 단어의 발음은 거의 바뀌지 않으면서 억양만 달라지는 것을 확인할 수 있었다. 이는 λ가 너무 작으면 학습 초기에 쿨벡-라이블러 발산이 너무 빠르게 수렴되어 생성 모델이 스타일에 대한 정보를 충분히 배울 수 없기 때문인 것으로 사료된다. 대조적으로, λ가 너무 크면 입력 음성과 완전히 다른 문장을 발화하는 것으로 확인되었다. 이것은 사전 분포와 사후 분포의 쿨벡-라이블러 발산값의 수렴 실패로 인한

것이다. 실험을 바탕으로, 적절한 λ를 찾는 것은 성공적인 학습에 매우 중요한 역할을 했다. 경험적으로 $\lambda = 2.0$ 및 $K = 16$인 모델이 가장 좋은 성능을 보였다.

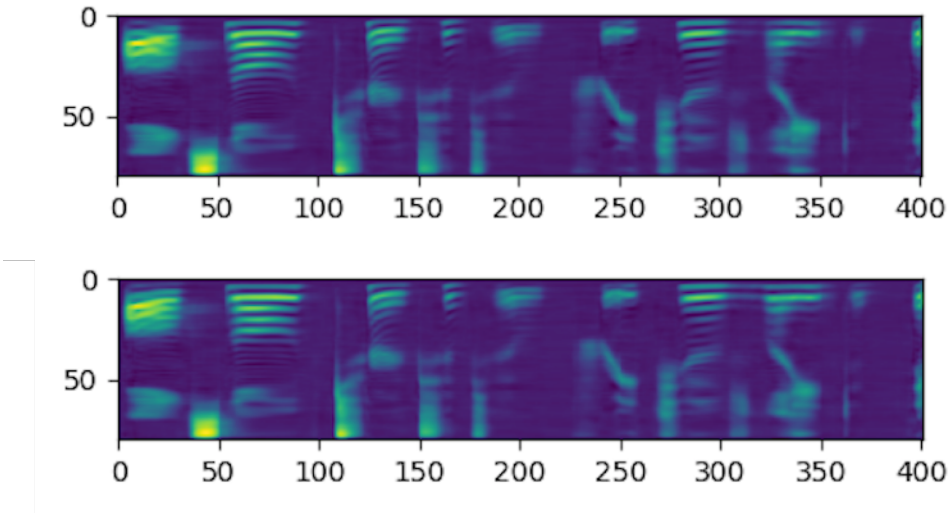

Figure 4.8: $\lambda = 1$일 때 멜-스펙트로그램 변화

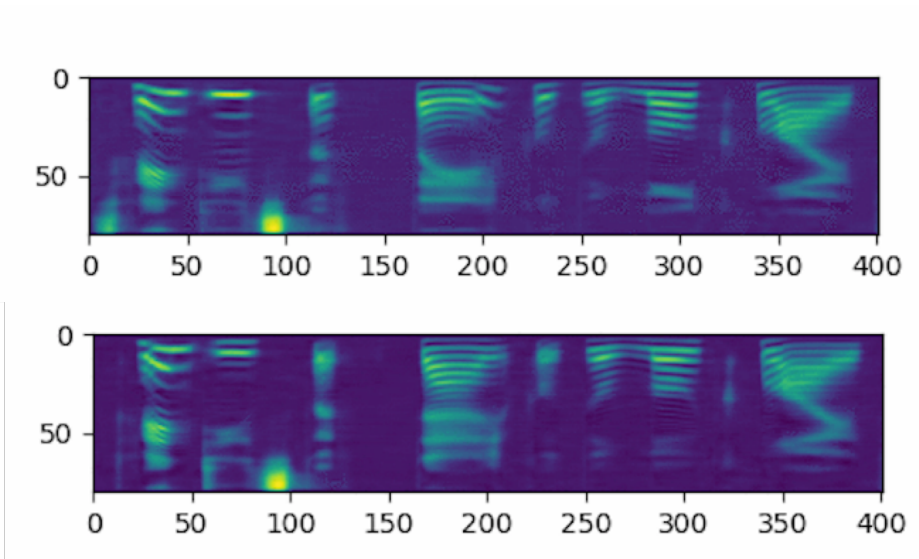

Figure 4.9: $\lambda = 2$일 때 멜-스펙트로그램 변화

제 5 장 결론

5.1 결론

음성 변환은 음성 합성과 동시에 기계 학습 분야에서 중요한 이슈로 떠올랐다. 음성 변환의 가장 대표적인 방법은 음성-문자 변환 후 문자-음성 변환의 2단계로 나뉘게 된다. 본 연구에서는 문자-음성 변환 과정에 조건부 변이 자동 인코더를 사용하여 확률적 정보를 추가하고, 이를 통해 다양한 억양을 가지는 발화를 생성하는 새로운 음성 변환 모델을 제시하였다. 본 모델은 다양한 원천 화자의 목소리를 한 명의 목표 화자의 목소리로 변환하는 모델이며, 다양한 억양이라 함은 목표 화자가 동일한 문장을 읽으며 발화할 때 매 샘플마다 다른 억양을 가지는 발화를 생성하는 것을 의미한다.

본 논문 이전에도 변분 자동 인코더를 사용한 음성 변환 연구가 존재했지만, 이 연구들은 언어 특성을 잠재 공간에 매핑하였기 때문에 억양에 대한 확률적 정보를 음성 변환에 활용할 수 없었다. 이에 비해 본 논문은 언어 기능을 사전 교육하고 고정시킨 뒤 잠재공간에 억양에 대한 정보를 매핑함으로써 억양의 다양성을 음성 변환에 적용할 수 있었다. 또한 다양한 억양을 가지는 음성을 생성하는 것뿐만 아니라 기존의 결정론적인 모델에 비해 보다 우수한 MOS 성능을 얻을 수 있었다.

다양한 억양을 가지는 발화를 생성하는 것을 확인한 후에 잠재 변수를 선형적으로 보간하는 실험을 진행하였다. 보간 실험은 억양이 어떤 규칙을 가지고 선형적으로 잠재 공간에 매핑되어 있는지 확인하는 작업이므로 그 의의가 크다. 보간 실험의 결과로 잠재 공간의 변화가 억양과 선형적인 관계가 있다는 것을 찾기는 힘들었다. 또한 프리빗 알고리즘은 사용한 학습은 경험적으로 음성의 억양 변화에 중요한 영향을 미친다는

것을 보여주었다. 본 논문은 또한 더 복잡한 사후확률분포를 만들기 위해 역 자기회귀 성 유동을 사용했다. 결과적으로 역 자기회귀성 유동을 적용한 모델은 전반적으로보다 다양한 억양을 가지는 음성을 생성했다. 이 외에도 손실함수로 멜-스펙트로그램과 선형 스펙트로그램을 사용했을 때의 차이, 그리고 다양한 하이퍼-파라미터를 조정하면서 다양한 절제 연구를 진행하였다.

이 연구는 음성 변환에 국한되지 않고 기존의 문자열에서 음성을 생성하는 TTS 분야에서도 활용될 수 있기 때문에 그 활용도가 높다. 특히 같은 문장을 읽어도 감정에 따라 문장의 발화가 달라지듯이 이 연구를 활용하면 다양한 억양을 가지는 발화를 문자열에서 생성해낼 수 있을 것이다.

5.2 향후 발전 방향

이 연구는 두가지 한계점이 존재한다. 첫번째는 생성된 샘플의 음질이 완전하지 않다는 것이다. 샘플의 음질을 향상시키기 위해서는 음성 분류기의 성능을 높이거나 음성 합성기에서 주의 구조를 이용하는 방법을 시도해볼 수 있다. 음소 분류기는 모델의 구조를 자기회귀성 구조로 바꾸어 성능을 높일 수 있을 것으로 에상된다. 음소 분류기의 특성 상 프레임 별 음소를 예측하게 되는데, 이전의 예측된 음소를 바탕으로 프레임과 함께 다음 음소를 예측하는 구조를 설계한다면 음소는 앞뒤 음소간의 관계가 중요하기 때문에 예측의 정확도가 높아질 것으로 예상된다. 음성 합성기의 경우 예측된 음소들 중 어느 음소를 집중적으로 볼지에 대한 주의 구조를 설계한다면 더욱 성능을 높일 수 있을 것으로 예상된다. 기존에 주의 구조를 사용한 연구들이 이미 좋은 성과를 내고 있고, 음성 합성의 분야에서도 주의 구조는 필수적으로 활용되고 있기 때문에 주의 구조의 장점을 활용한다면 합성된 음성의 음질을 높이는데에 좋은 영향을 끼칠 것으로 예상된다.

두번째로, 이 연구는 다양한 억양을 가지는 음성을 생성하는 모델을 입증했지만, 원하는 억양을 추출해내기 위해 잠재 공간을 제어하지는 못한다는 것을 보간 실험을 통해 알아내었고, 이런 측면에서 한계점이 존재한다. 잠재 공간을 제어하지 못하면 다양한 억양에 대한 제어가 불가능해지며, 이렇게 되면 다양한 억양을 가지는 음성을 생성하더라도 원하는 억양으로 생성하지 못하게 되기 때문에 실제로 서비스에 적용하거나 유용하게 사용하기 위해서는 향후 연구가 필수적이다. 잠재공간을 제어하기 위해서는 사후 확률분포가 잘 나뉘어져있고, 사전확률분포 역시 더 복잡한 분포를 사용하는 연구를 생각해볼 수 있을 것이다. 향후 보다 복잡하고 제어 가능한 사전 확률 네트워크를 설계함으로써 이 문제를 해결할 것으로 기대한다.

다양한 억양을 가지는 음성을 합성하는 연구의 의의는 최종적으로 사람들이 어떤

상황에서 어떤 억양으로 발화를 하는지, 각 상황에 맞는 억양을 가지는 발화의 생성이다. 향후에 감정 등의 레이블을 조건으로 추가하여 네트워크를 학습하고 감정에 따라 다른 억양을 가지는 음성을 생성하는 연구도 진행될 수 있을 것이다.

참고 문헌

[1] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[2] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

[3] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

[4] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135*, 2017.

[5] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad. Voice conversion using artificial neural networks. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3893–3896. IEEE, 2009.

[6] Srinivas Desai, Alan W Black, B Yegnanarayana, and Kishore Prahallad. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):954–964, 2010.

[7] Zhi-Zheng Wu, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li. Text-independent f0 transformation with non-parallel data for voice conversion. In *Eleventh annual conference of the international speech communication association*, 2010.

[8] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.

[9] Dimitrios Rentzos, Saeed Vaseghi, Qin Yan, and Ching-Hsiang Ho. Voice conversion through transformation of spectral and intonation features. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–21. IEEE, 2004.

[10] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, pages 1–6. IEEE, 2016.

[11] Chen Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849*, 2017.

[12] Sercan O Arık, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *arXiv preprint arXiv:1705.08947*, 2017.

[13] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433*, 2017.

[14] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[15] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[16] Hiroyuki Miyoshi, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Voice conversion using sequence-to-sequence learning of context posterior probabilities. *arXiv preprint arXiv:1704.02360*, 2017.

[17] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6309–6318, 2017.

[18] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *arXiv preprint arXiv:1712.05884*, 2017.

[19] Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Jonathan Raiman, Shubho

Sengupta, et al. Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*, 2017.

[20] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *arXiv preprint arXiv:1710.07654*, 2017.

[21] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. Voice synthesis for in-the-wild speakers via a phonological loop. *arXiv preprint arXiv:1707.06588*, 2017.

[22] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[23] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[24] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE, 2015.

[26] John Kominek and Alan W Black. The cmu arctic speech databases. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *CoRR, abs/1703.06868*, 2017.

Abstract

Voice Conversion with Diverse Intonation using Conditional Variational Auto-Encoder

Soobin Suh
Department of Industrial Engineering
The Graduate School
Seoul National University

Voice conversion is a task of synthesizing an utterance with target speaker's voice while maintaining linguistic information of the source utterance. While a speaker can produce varying utterances from a single script with different intonations, conventional voice conversion models were limited to producing only one result per source input. To overcome this limitation, we propose a novel approach for voice conversion with diverse intonations using conditional variational autoencoder (CVAE).

Experiments have shown that the speaker's style feature can be mapped into a latent space with Gaussian distribution. We have also been able to convert voices with more diverse intonation by making the posterior of the latent space more complex with inverse autoregressive flow (IAF). As a result, the converted voice not only has a diversity of intonations, but also has better sound quality than the model without CVAE.

Keywords: Voice Conversion, Variational Auto-Encoder, Intonations

Student Number: 2016-24163